Bilevel Learning of the Group Lasso Structure

Jordan Frecon\textsuperscript{1}, Saverio Salzo\textsuperscript{1}, Massimiliano Pontil\textsuperscript{1,2}

\textsuperscript{1} CSML - Istituto Italiano di Tecnologia
\textsuperscript{2} Dept of Computer Science - University College London

Thirty-second Conference on Neural Information Processing Systems, Montreal, Canada
Linear Regression and Group Sparsity

**Problem:** Predict $y \in \mathbb{R}^N$ from $X \in \mathbb{R}^{N \times P}$

**Linear Regression:** Find $w \in \mathbb{R}^P$ such that

In many applications, few groups are relevant to predict $y \Rightarrow$ Group Sparse $w$

- Predict psychiatric disorder from activities in regions of the brain
- Predict protein functions from their molecular composition
Group Lasso

Given $\lambda > 0$ and a group-structure $\{G_1, \ldots, G_L\}$, find

$$\hat{w} \in \arg \min_{w \in \mathbb{R}^p} \frac{1}{2} \|y - Xw\|^2 + \lambda \sum_{l=1}^{L} \|w_{G_l}\|_2,$$

**Limitation:** The group-structure $\{G_1, \ldots, G_L\}$ may be unknown.
Setting: \( T \) Group Lasso problems with shared group-structure

\[
(\forall t \in \{1, \ldots, T\}) \quad \hat{w}_t(\theta) \in \arg\min_{w_t \in \mathbb{R}^P} \frac{1}{2} \|y_t - X_t w_t\|^2 + \lambda \sum_{l=1}^{L} \|w_t \odot \theta_l\|_2,
\]

Goal: Estimation of the optimal group-structure \( \theta^* \)
A Bilevel Programming Approach

**Upper-level Problem:**

\[
\begin{align*}
\text{minimize } & \ U(\theta) := \sum_{t=1}^{T} \mathcal{E}_t(\hat{w}_t(\theta)) & \quad (\text{e.g., validation error}) \\
\text{subject to } & \ \theta_1, \ldots, \theta_L \in \Theta \\
\end{align*}
\]

where \( \hat{w}(\theta) = [\hat{w}_1(\theta) \cdots \hat{w}_T(\theta)] \) solves

**Lower-level Problem:** (\( T \) Group Lasso problems)

\[
\begin{align*}
\text{minimize } & \ L(w, \theta) := \sum_{t=1}^{T} \left( \frac{1}{2} \| y_t - X_t w_t \|_2^2 + \lambda \sum_{l=1}^{L} \| \theta_l \odot w_t \|_2 \right) \\
\end{align*}
\]

**Difficulties:**

- \( \hat{w}(\theta) \) not available in closed form
- \( \theta \mapsto \hat{w}(\theta) \) is nonsmooth \( \Rightarrow U \) is nonsmooth
Approximate Bilevel Problem

Upper-level Problem:

\[
\begin{align*}
\text{minimize } & \quad \mathcal{U}_K(\theta) := \sum_{t=1}^{T} \mathcal{E}_t(w_t^{(K)}(\theta)) \\
\text{where } & \quad w_t^{(K)}(\theta) \to \hat{w}_t(\theta)
\end{align*}
\]

Dual Algorithm:

\[
\begin{align*}
u^{(0)}(\theta) & \text{ chosen arbitrarily} \\
\text{for } k = 0, 1, \ldots, K - 1 & \\
\quad u^{(k+1)}(\theta) &= A(u^{(k)}(\theta), \theta) \quad \text{dual update} \\
\left[ w_1^{(K)}(\theta) \cdots w_T^{(K)}(\theta) \right] &= B(u^{(K)}(\theta), \theta) \quad \text{primal dual relationship}
\end{align*}
\]

Goals:

- Find \(A\) and \(B\) smooth \(\Rightarrow w^{(K)}\) is smooth \(\Rightarrow \mathcal{U}_K\) is smooth
- Prove that the approximate bilevel scheme converges
Contributions

- Bilevel Framework for Estimating the Group Lasso Structure

- Design of a Dual Forward-Backward Algorithm with Bregman Distances such that
  1. $\mathcal{A}$ and $\mathcal{B}$ are smooth $\Rightarrow U_K$ is smooth
  2. $\begin{cases} \min U_K \rightarrow \min U \\ \arg\min U_K \rightarrow \arg\min U \end{cases}$

Implementation of proxSAGA algorithm: nonconvex stochastic variant of

$$
\theta^{(q+1)} = P_\Theta (\theta^{(q)} - \gamma \nabla U_K (\theta^{(q)}))
$$
Numerical Experiment

**Setting:** $T = 500$ tasks, $N = 25$ noisy observations, $P = 50$ features. Estimate and group the features into, at most, $L = 10$ groups.
Thank You

Our poster AB #92 will be presented in Room 210 & 230 at 5pm